

Measuring Semantic Relatedness using Mined Semantic Analysis

Walid Shalaby, Wlodek Zadrozny
Computer Science Department
UNC Charlotte
{wshalaby, wzadrozny}@uncc.edu

ABSTRACT

Mined Semantic Analysis (*MSA*) is a novel concept space model which employs unsupervised learning to generate semantic representations of text. *MSA* represents textual structures (terms, phrases, documents) as a bag-of-concepts where concepts are derived from concept rich encyclopedic corpora. Traditional concept space models exploit only target corpus content to construct the concept space. *MSA*, alternatively, uncovers implicit relations between concepts by mining for their associations (e.g., mining Wikipedia's "See also" link graph). We evaluate *MSA*'s performance on benchmark data sets for measuring lexical semantic relatedness. Empirical results show competitive performance of *MSA* compared to prior state-of-the-art methods. Additionally, we introduce the first analytical study to examine statistical significance of results reported by different semantic relatedness methods. Our study shows that, the nuances of results across top performing methods could be statistically insignificant. The study positions *MSA* as one of state-of-the-art methods for measuring semantic relatedness.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Algorithms, Measurement

Keywords

Semantic relatedness, concept space models, bag-of-concepts, association rule mining

1. INTRODUCTION

For decades semantic analysis of textual content has gained enormous attention within the Natural Language Processing (*NLP*) community as a means for automating language understanding. To this end, evaluating lexical semantic similarity/relatedness has attracted many researchers as an enabler mechanism for many text understanding tasks. Semantic relatedness is a knowledge intensive task

as it requires huge amount of world knowledge to accomplish its goal [Hassan and Mihalcea, 2011].

Although semantic similarity and relatedness are often used interchangeably in the literature, they do not represent the same task [Budanitsky and Hirst, 2006]. Evaluating genuine similarity is, and should be, concerned with measuring the similarity or resemblance in meanings and hence focuses on the synonymy relations (e.g., *smart,intelligent*). Relatedness, on the other hand, is more general and covers broader scope as it focuses on other relations such as antonymy (*old,new*), hypernymy (*cock,bird*), and other functional associations (*money,bank*).

Semantic relatedness has many applications in *NLP* and Information Retrieval (*IR*) for addressing problems such as word sense disambiguation, paraphrasing, text categorization, dimensionality reduction, and others. Most semantic relatedness methods are inspired by the distributional hypothesis [Harris, 1954] which emphasizes the idea that similar words tend to appear in similar contexts and thus have similar contextual distributions. Those methods often develop a distributional semantics model which represents each linguistic term as a vector derived from contextual information of that term in a large corpus of text or knowledge base. [Baroni et al., 2014, Turney et al., 2010, Hassan and Mihalcea, 2011, Gabrilovich and Markovitch, 2007, Landauer et al., 1997]. After constructing such distributional vectors, relatedness is calculated using an appropriate vector similarity measure (e.g., cosine similarity).

In this paper we propose Mined Semantic Analysis (*MSA*), a novel concept space model for semantic analysis using unsupervised data mining techniques. *MSA* represents textual structures (terms, phrases, documents) as a bag-of-concepts. Unlike other concept space models which look for direct associations between concepts and terms through statistical co-occurrence [Camacho-Collados et al., 2015, Hassan and Mihalcea, 2011, Gabrilovich and Markovitch, 2007], *MSA* discovers implicit concept-concept associations using rule mining [Agrawal et al., 1993]. *MSA* uses these associations subsequently to enrich term's concept space with latent concepts.

MSA utilizes a search index created using concept rich corpora (e.g., *Wikipedia*). The concept space of a given term is constructed through two phases. First, an initial set of candidate concepts is retrieved from the index. Second, the candidates set is augmented with other related concepts using discovered concept-concept association rules. Following this strategy, *MSA* identifies not only concepts directly related to a given text but also other latent concepts associated implicitly with them.

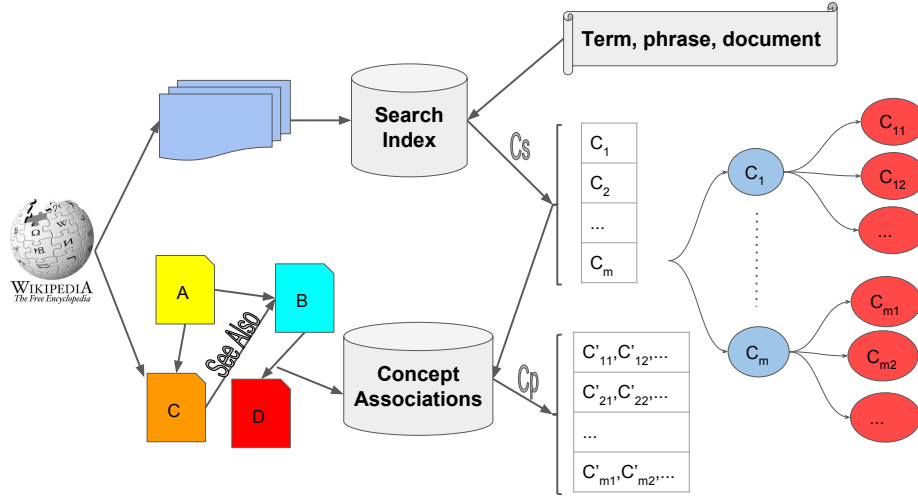


Figure 1: MSA generates the concept space of a given textual structure through: 1) explicit concept retrieval from the index (top); and 2) concept expansion from the concept-concept associations repository (bottom).

The contributions of this paper are threefold: First, we introduce a novel concept space model for semantic analysis which augments explicit semantics with conceptual associations through data mining techniques. Second, we demonstrate the effectiveness of this method for evaluating semantic relatedness on benchmark data sets. Third, we present the first analytical study to examine statistical significance of results reported by different semantic relatedness methods. Our study shows that, the nuances of results across top performing methods could be statistically insignificant. The study positions *MSA* as one of state-of-the-art methods for measuring semantic relatedness.

2. RELATED WORK

Several semantic representation models have been proposed in the literature. Some of them utilize textual corpora from which world knowledge is acquired and used to represent textual structures as high-dimensional "meaning" vectors. As pointed out by [Baroni et al., 2014], those vectors are either estimated by means of statistical modeling such as *LSA* [Landauer et al., 1997] and *LDA* [Blei et al., 2003], or more recently through neural network based representations such as *CW* [Collobert and Weston, 2008], *Word2Vec* [Mikolov et al., 2013], and *GloVe* [Pennington et al., 2014].

Knowledge-based models were also proposed for measuring semantic relatedness [Jarmasz and Szpakowicz, 2004, Budanitsky and Hirst, 2006, Zesch et al., 2008, Pilehvar and Navigli, 2015]. Those models utilize dictionaries such as *Wordnet* [Fellbaum, 1998] and *Wiktionary*, and use explicit word relations to infer semantic relatedness. Hybrid models which incorporate knowledge from corpora and dictionaries were also used to evaluate semantic relatedness [Agirre et al., 2009, Banjade et al., 2015, Camacho-Collados et al., 2015].

Explicit concept space models such as *ESA* [Gabrilovich and Markovitch, 2007], *SSA* [Hassan and Mihalcea, 2011], and *NASARI* [Camacho-Collados et al., 2015] construct bag-of-concepts (*BOC*) vectors to represent textual structures using concepts in encyclopedic knowledge source such as Wikipedia. Those *BOC* embeddings capture the main topics of the given text and therefore are useful for under-

standing its semantics.

The *BOC* representations have proven efficacy for semantic analysis of textual data especially short texts where contextual information is missing or insufficient. For example, measuring lexical semantic similarity/relatedness [Gabrilovich and Markovitch, 2007], text categorization [Song and Roth, 2014], search and relevancy ranking [Egozi et al., 2011], and others. Semantic relatedness models typically employ those semantic vectors to measure relatedness using appropriate similarity measure between the vectors.

A closely related method to *MSA* is Explicit Semantic Analysis (*ESA*) [Gabrilovich and Markovitch, 2007]. *ESA* constructs the concept space of a term by searching an inverted index of term-concept co-occurrences. *ESA* is mostly the traditional vector space model applied to *Wikipedia* articles. *ESA* is effective in retrieving concepts which explicitly mention the target search terms in their content. However, it fails to identify other latent concepts which do not contain the search terms. *MSA* bridges this gap by mining for concept-concept associations and thus augmenting the concept space identified by *ESA* with more relevant concepts.

Salient Semantic Analysis (*SSA*) was proposed by [Hassan and Mihalcea, 2011] and uses *Wikipedia* concepts to build semantic profiles of words. *SSA* is more conservative than *ESA* as it defines word meaning by its immediate context and therefore might yield concepts of higher relevancy. However, it is still limited to surface semantic analysis because it, like *ESA*, utilizes only direct associations between words and concepts and fails to capture other latent concepts not directly co-occurring with corpus words in the same context.

[Radinsky et al., 2011] proposed Temporal Semantic Analysis (*TSA*) which works by extending *ESA*'s concept space to include temporal usage patterns of discovered concepts. Both *MSA* and *TSA* share a common goal; they try to complement the concept space with information that uncovers implicit concept associations. However, they follow totally different methodologies for achieving that goal. *TSA* exploits temporal dynamics of concept usage, while *MSA* ex-

Explicit Concepts	Latent Concepts
Parse tree	Universal Networking Language
Temporal annotation	Translation memory
Morphological dictionary	Systemic functional linguistics
Textalytics	Semantic relatedness
Bracketing	Quantitative linguistics
Lemmatization	Natural language processing
Indigenous Tweets	Internet linguistics
Statistical semantics	Grammar induction
Treebank	Dialog systems
Light verb	Computational semiotics

Table 1: The concept space of "Computational Linguistics"

exploits mining the semantic space of each concept as expressed in its associations with other concepts.

Another closely related model is Latent Semantic Analysis (*LSA*) [Deerwester et al., 1990, Landauer et al., 1997]. *LSA* is a statistical model that was originally proposed to solve the vocabulary mismatch problem in information retrieval. *LSA* first builds a term-document co-occurrence matrix from textual corpus and then maps that matrix into a new space using singular-value decomposition. In that semantic space terms and documents that have similar meaning will be placed close to one another. Though its effectiveness, *LSA* has been known to be hard to explain because it is difficult to map the computed space dimensions into meaningful concepts. *MSA*, alternatively, generates explicit conceptual mappings that are interpretable by humans making it more intuitive than *LSA*.

3. MINED SEMANTIC ANALYSIS

We call our approach Mined Semantic Analysis (*MSA*) as it utilizes data mining techniques in order to discover the concept space of textual structures. The motivation behind our approach is to mitigate a notable gap in prior concept space models which are limited to direct associations between words and concepts. Therefore those models lack the ability to transfer the association relation to other latent concepts which contribute to the meaning of these words.

Figure 1 shows *MSA*'s architecture. In a nutshell, *MSA* generates the concept space of a given text by utilizing two repositories created offline: 1) a search index of *Wikipedia* articles, and 2) a concept-concept associations repository created by mining the "See also" link graph of *Wikipedia* concepts (articles). First, the explicit concept space is constructed by retrieving concepts (titles of articles) explicitly mentioning the given text. Second, latent concepts associated with each of the explicit concepts are retrieved from the associations repository and used to augment the concept space.

To demonstrate our approach, we provide an example of exploring the concept space of "*Computational Linguistics*" (Table 1). Column 1 shows the explicit concepts retrieved by searching *Wikipedia*¹. Column 2 shows the same explicit concepts in column 1 enriched by implicit concepts. As we can notice, those implicit concepts could augment the explicit concept space by more related concepts which contribute to understanding "*Computational Linguistics*". It is worth mentioning that not all implicit concepts are equally relevant, therefore we also propose an automated mechanism for ranking those concepts in a way that reflects their relatedness to the original search term.

¹We search *Wikipedia* using a term-concept inverted index and limit the search space to articles with min. length of 2k and max. title length of 3 words.

3.1 The Search Index

MSA starts constructing the concept space of term(s) by searching for an initial set of candidate explicit concepts. For this purpose, we build a search index of a concept rich corpus such as *Wikipedia* where each article represents a concept. This is similar to the idea of the inverted index introduced in *ESA* [Gabrilovich and Markovitch, 2007]. We build the index using *Apache Lucene*², an open-source indexing and search engine. For each article we index the title, content, length, and the "See also" section.

During search we use some parameters to tune the search space. Specifically, we define the following parameters to provide more control over search:

Article Length (*L*): minimum length of *Wikipedia* article in characters excluding sections like "*References*", "*See also*", "*Categories*", ...etc.

Number of Concepts (*M*): maximum number of concepts (articles) to retrieve as initial candidate concepts.

Title Length (τ): this threshold is important for pruning all articles that have long irrelevant titles. It represents the maximum number of words in the title, for example, if $\tau=3$, then all articles with more than three words in title will be pruned.

3.2 Association Rules Mining

In order to discover the implicit concepts, we employ rule mining [Agrawal et al., 1993] to learn implicit relations between concepts using *Wikipedia*'s "See also" link graph.

Formally, given a set of concepts $C = \{c_1, c_2, \dots, c_N\}$ of size N (i.e., all *Wikipedia* articles). We build a dictionary of transactions $T = \{t_1, t_2, t_3, \dots, t_M\}$ of size M such that $M \leq N$. Each transaction t in T contains a subset of concepts in C . t is constructed from each article in *Wikipedia* that contains at least one entry in its "See also" section. For example, if an article representing concept c_1 with entries in its "See also" section referring to concepts $\{c_2, c_3, \dots, c_n\}$, a transaction $t = \{c_1, c_2, c_3, \dots, c_n\}$ will be constructed and added to T . A set of rules R is then created by mining T . Each rule r in R is defined as in equation 1:

$$r(s, f) = \{(X \Rightarrow Y) : X, Y \subseteq C \text{ and } X \cap Y = \emptyset\} \quad (1)$$

Both X and Y are subsets of concepts in C . X are called the antecedents of r and Y are called the consequences. Rule r is parameterized by two parameters: 1) Support (s) which indicates how many times both X and Y appeared together in T , and 2) Confidence (f) which is s divided by number of times X appeared in T .

After learning R , we end up having concept(s)-concept(s) associations. Using such rules, we can determine the strength of those associations based on s and f .

As the number of rules grows exponentially with the number of concepts, we define the following parameters to provide more fine grained control on participating rules during explicit concept expansion:

Consequences Size ($|Y|$): number of concepts in rule consequences (right hand side).

²<http://lucene.apache.org/core/>

Minimum Support (ϵ): minimum rule support. It defines the minimum strength of the association between rule concepts. For example, if $\epsilon = 2$, then all rules whose support $s \geq 2$ will be considered during concept expansion.

Minimum Confidence (v): this threshold defines the minimum strength of the association between rule concepts compared to other rules with same antecedents. For example, if $v = 0.5$, then all rules whose confidence $f \geq 0.5$ will be considered during concept expansion. In other words, consequent concept(s) must have appeared in at least 50% of the times antecedent concept(s) appeared in T .

3.3 Constructing the Concept Space

Given a set of concepts C of size N , MSA constructs the bag-of-concepts vector C_t of term(s) t through two phases: *Search* and *Expansion*. In the search phase, t is represented as a search query and is searched for in the *Wikipedia* search index. This returns a weighted set of articles that best matches t based on the vector space model. We call the set of concepts representing those articles C_s and is represented as in equation 2:

$$C_s = \{(c_i, w_i) : c_i \in C \text{ and } i \leq N\} \\ \text{subject to :} \\ |title(c_i)| \leq \tau, length(c_i) \geq L, |C_s| \leq M \quad (2)$$

Note that we search all articles whose content length and title n-grams meet the thresholds L and τ respectively. The weight of c_i is denoted by w_i and represents the match score between t and c_i as returned by the search engine.

In the expansion phase, we use inferred association rules to expand each concept c in C_s by looking for its associated set of concepts in R . Formally, the expansion set of concepts C_p is obtained as in equation 3:

$$C_p = \bigcup_{c \in C_s, c' \in C} \{(c', w) : \exists r(s, f) = c \Rightarrow c'\} \\ \text{subject to : } |c'| = |Y|, s \geq \epsilon, f \geq v \quad (3)$$

Note that we add all the concepts that are implied by c where this implication meets the support and confidence thresholds (ϵ, v) respectively. The weight of c' is denoted by w ; currently we use simple weight propagation mechanism where all concepts implied by c inherit the same weight as c .

Finally, all the concepts from search and expansion phases are merged to construct the concept vector C_t of term(s) t as in equation 4:

$$C_t = C_s \cup C_p \quad (4)$$

3.4 Relatedness Scoring

In order to calculate the relatedness score between a term pair (t_1, t_2) , we first sparsify their concept vectors (C_{t_1}, C_{t_2}) to have same length. We then apply the traditional cosine similarity measure on their respective weight vectors (W_{t_1}, W_{t_2}) as in equation 5:

$$Rel_{cos}(t_1, t_2) = \frac{W_{t_1} \cdot W_{t_2}}{\|W_{t_1}\| \|W_{t_2}\|} \quad (5)$$

Similar to [Hassan and Mihalcea, 2011], we include a normalization factor λ as the cosine measure gives low scores for highly related terms due to their concept vectors sparsity. Other approaches for dealing with vector sparsity worth exploring in the future [Song

and Roth, 2015]. Using λ , the final relatedness score will be adjusted as in equation 6:

$$Rel(t_1, t_2) = \begin{cases} 1 & Rel_{cos}(t_1, t_2) \geq \lambda \\ \frac{Rel_{cos}(t_1, t_2)}{\lambda} & Rel_{cos}(t_1, t_2) < \lambda \end{cases} \quad (6)$$

4. EXPERIMENTS AND RESULTS

4.1 Data Sets

We evaluate MSA 's performance on benchmark data sets for measuring lexical semantic relatedness. Each data set is a collection of word pairs along with human judged similarity/relatedness score for each pair.

RG: a similarity data set created by [Rubenstein and Goodenough, 1965]. It contains 65 noun pairs³. Similarity judgments of each pair were conducted by 51 subjects. Judgments range from 0 (very unrelated) to 4 (very related). [Pilehvar and Navigli, 2015] reported highest performance on this data set by creating a semantic network from *Wiktionary*.

MC: a similarity data set created by [Miller and Charles, 1991]. It contains 30 noun pairs⁴ taken from *RG* data set. Similarity judgments were done by 38 subjects at the same scale as *RG*. [Camacho-Collados et al., 2015] reports the highest performance on *MC* by integrating knowledge from *Wikipedia* and *Wordnet*.

WS: a relatedness data set created by [Finkelstein et al., 2001] and contains 353 word pairs⁵. Relatedness score for each pair was judged by 13-16 annotators ranging from 0 (totally unrelated) to 10 (very related). Annotators were not instructed to differentiate between similarity and relatedness. [Halawi et al., 2012] reports the highest performance on *WS* using a supervised model combined with constraints of known related words.

WSS & WSR: [Agirre et al., 2009] manually split *WS* data set into two subsets to separate between similar and related pairs⁶. *WSS* contains 203 similar word pairs. *WSR* contains 252 related word pairs. [Baroni et al., 2014] reports the highest performance on both data sets using the popular neural network based model *Word2Vec*⁷ proposed by [Mikolov et al., 2013].

MEN: a relatedness data set created by [Bruni et al., 2014]⁸. We use the test subset of this data set which contains 1000 pairs. Relatedness scores range from 0 (totally unrelated) to 50 (totally related). [Baroni et al., 2014] reports the highest performance on this collection using *Word2Vec*⁹.

4.2 Experimental Setup

We followed experimental setup similar to [Baroni et al., 2014]. Basically, we implemented two sets of experiments. First, we perform a grid search over MSA 's parameter space to obtain the maximum performing combination of parameters on each data set. Second, we evaluate MSA in a more realistic settings where we use one of the data sets as a development set for tuning MSA 's parameters and then use tuned parameters to evaluate MSA 's performance on

³<http://www.cs.cmu.edu/~mfaruqui/word-sim/EN-RG-65.txt>

⁴<http://www.cs.cmu.edu/~mfaruqui/word-sim/EN-MC-30.txt>

⁵<http://alfonseca.org/eng/research/wordsim353.html>

⁶<http://alfonseca.org/eng/research/wordsim353.html>

⁷<https://code.google.com/p/word2vec/>

⁸<http://clic.cimec.unitn.it/~elia.bruni/MEN.html>

⁹<https://code.google.com/p/word2vec/>

	<i>MC</i>	<i>RG</i>	<i>WSS</i>	<i>WSR</i>	<i>WS</i>
<i>LSA</i> [*]	0.73	0.64	–	–	0.56
<i>ESA</i> [◊]	0.74	0.72	0.45	–	0.49 [*]
<i>SSA</i> _s [*]	0.87	0.85	–	–	0.62
<i>SSA</i> _c [*]	0.88	0.86	–	–	0.59
<i>ADW</i> [◊]	0.79	0.91	0.72	–	–
<i>NASARI</i> [◊]	0.91	0.91	0.74	–	–
<i>Word2Vec</i> [▷]	0.82	0.84	0.76	0.65	0.68
<i>MSA</i>	0.91	0.87	0.77	0.66	0.69

Table 2: *MSA*’s Pearson (r) scores on benchmark data sets vs. other techniques. (^{*}) from [Hassan and Mihalcea, 2011], ([▷]) from [Baroni et al., 2014] predict vectors, ([◊]) from [Camacho-Collados et al., 2015].

the other data sets. In both sets of experiments, we set $|Y| = 1$ and $v = 0.0$.

We built the search index using *Wikipedia* dump of March 2015¹⁰. The total uncompressed XML dump size was about 52GB representing about 7 million articles. We extracted the articles using a modified version of Wikipedia Extractor¹¹. Our version¹² extracts articles plain text discarding images and tables. We also discard *References* and *External links* sections (if any). We pruned both articles not under the main namespace and pruned all redirect pages as well. Eventually, our index contained about 4.8 million documents in total.

4.3 Evaluation

We report the results by measuring correlation between *MSA*’s computed relatedness scores and the gold standard provided by human judgments. As in prior studies, we report both Pearson correlation (r) [Hill and Lewicki, 2007] and Spearman rank-order correlation (ρ) [Zwillinger and Kokoska, 1999].

We compare our results with those obtained from three types of semantic representation models. First, statistical co-occurrence models like *LSA* [Landauer et al., 1997], *CW* and *BOW* [Agirre et al., 2009], and *ADW* [Pilehvar and Navigli, 2015]. Second, neural network models like Collobert and Weston (*CW*) vectors [Collobert and Weston, 2008], *Word2Vec* [Baroni et al., 2014], and *GloVe* [Pennington et al., 2014]. Third, explicit semantics models like *ESA* [Gabrilovich and Markovitch, 2007], *SSA* [Hassan and Mihalcea, 2011], and *NASARI* [Camacho-Collados et al., 2015].

4.4 Results

We report correlation scores of *MSA* compared to other models in Tables 2, 3, and 4. Some models do not report their correlation scores on all data sets, so we leave them blank. *MSA* (last row) represents scores obtained by using *WS* as a development set for tuning *MSA*’s parameters and evaluating performance on the other data sets using the tuned parameters. The parameter values obtained by tuning on *WS* were $L = 5k$, $M = 800$, $\tau = 2, 3$ for C_s , C_p respectively, and finally $\epsilon = 1$.

Table 2 shows Pearson correlation (r) of *MSA* on five benchmark data sets. It also report prior work results on same data sets. For

	<i>MC</i>	<i>RG</i>	<i>WSS</i>	<i>WSR</i>	<i>WS</i>
<i>LSA</i> [*]	0.66	0.61	–	–	0.58
<i>ESA</i> [‡]	0.70	0.75	0.53	–	0.75
<i>SSA</i> _s [*]	0.81	0.83	–	–	0.63
<i>SSA</i> _c [*]	0.84	0.83	–	–	0.60
<i>CW</i> ^Υ	–	0.89	0.77	0.46	0.60
<i>BOW</i> ^Υ	–	0.81	0.70	0.62	0.65
<i>NASARI</i> [§]	0.80	0.78	0.73	–	–
<i>ADW</i> [◊]	0.90 ¹⁴	0.92	0.75	–	–
<i>GloVe</i> ^Ψ	0.84	0.83	–	–	0.76
<i>Word2Vec</i> [▷]	0.82 ¹⁵	0.84	0.76	0.64	0.71
<i>MSA</i>	0.87	0.86	0.77	0.71	0.73

Table 3: *MSA*’s Spearman (ρ) scores on benchmark data sets vs. other techniques. (^{*}) from [Hassan and Mihalcea, 2011], ([‡]) from [Pilehvar and Navigli, 2015, Hassan and Mihalcea, 2011], (^Υ) from [Agirre et al., 2009], ([§]) using pairwise similarities from [Camacho-Collados et al., 2015], ([◊]) from [Pilehvar and Navigli, 2015], (^Ψ) from [Pennington et al., 2014], ([▷]) from [Baroni et al., 2014]

	<i>MEN</i>
<i>Skipgram</i> [*]	0.44
<i>CW</i> [*]	0.60
<i>GloVe</i> [*]	0.71
<i>Word2Vec</i> [▷]	0.79
<i>MSA</i>	0.75

Table 4: *MSA*’s Spearman (ρ) scores on *MEN* data set vs. other techniques. (^{*}) from [Hill et al., 2014], ([▷]) from [Baroni et al., 2014]

Word2Vec, we obtained [Baroni et al., 2014] predict vectors¹³ and used them to calculate Pearson correlation scores. It is clear that, in absolute figures, *MSA* consistently gives the highest correlation scores on all data sets compared to other methods except on *RG* where *NASARI* and *ADW* [Camacho-Collados et al., 2015] performed better.

The best performance, in terms of Pearson correlation, obtained by performing grid search over *MSA*’s parameter space was 0.97 on *MC*, 0.90 on *RG*, 0.78 on *WSS*, 0.67 on *WSR*, and 0.69 on *WS*.

Table 3 shows *MSA*’s Spearman correlation scores compared to prior models on same data sets as in Table 2. As we can see, *MSA* gives highest scores on *WSS* and *WSR* data sets. It comes second on *MC*, third on *RG* and *WS*. We can notice that *MSA* consistently performed better than the popular *Word2Vec* model on all data sets. *MSA*’s latent concepts enrichment participated in performance gains compared to other explicit concept space models such as *ESA* and *SSA*.

The best performance, in terms of Spearman correlation, obtained by performing grid search of *MSA*’s parameter space was 0.95 on *MC*, 0.91 on *RG*, 0.78 on *WSS*, 0.72 on *WSR*, and 0.73 on *WS*.

Table 4 shows *MSA*’s Spearman correlation scores compared to

¹⁰<https://dumps.wikimedia.org/enwiki/20150304/>

¹¹http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

¹²<https://github.com/walid-shalaby/wikiextractor>

¹³<http://clit.cimec.unitn.it/composes/semantic-vectors.html>

¹⁴Pairwise similarity scores obtained by contacting authors of [Pilehvar and Navigli, 2015]

¹⁵Using <http://clit.cimec.unitn.it/composes/semantic-vectors.html>

	<i>MC</i>		<i>RG</i>		<i>WSS</i>		<i>WSR</i>		<i>WS</i>		<i>MEN</i>	
	ρ	<i>p</i> -value	ρ	<i>p</i> -value	ρ	<i>p</i> -value	ρ	<i>p</i> -value	ρ	<i>p</i> -value	ρ	<i>p</i> -value
	<i>MSA_t</i>											
<i>Word2Vec_t</i> [▷]	0.84	0.168	0.78	0.297	0.79	0.357	0.70	0.019	0.72	0.218	0.78	0.001
<i>NASARI</i> [*]	0.73	0.138	0.77	0.030	0.70	0.109	–	–	–	–	–	–
<i>ADW</i> [◊]	0.78	0.258	0.78	0.019	0.67	0.271	–	–	–	–	–	–
	<i>ADW[◊]</i>											
<i>Word2Vec_t</i> [▷]	0.80	0.058	0.81	0.003	0.68	0.5	–	–	–	–	–	–
<i>NASARI</i> [*]	0.82	0.025	0.80	0.0	0.76	0.256	–	–	–	–	–	–
	<i>Word2Vec_t</i> [▷]											
<i>NASARI</i> [*]	0.75	0.387	0.71	0.105	0.66	0.192	–	–	–	–	–	–

Table 5: Steiger’s Z significance test on the differences between Spearman correlations (ρ) using 1-tailed test and 0.05 statistical significance. (▷) using [Baroni et al., 2014] predict vectors, (*) using [Camacho-Collados et al., 2015] pairwise similarity scores, (◊) using [Pilehvar and Navigli, 2015] pairwise similarity scores.

other models (all are neural network models). As we can see, *MSA* comes second after *Word2Vec* giving higher correlation than *Skip-gram*, *CW*, and *GloVe*. Results on this data set prove that *MSA* is a very advantageous method for evaluating lexical semantic relatedness compared to the popular deep learning models. On another hand, *MSA*’s Pearson correlation score on *MEN* data set was 0.73.

We can notice from the results in Tables 2 and Table 3 that measuring semantic relatedness is more difficult than measuring semantic similarity. This is clear from the drop in correlation scores of the relatedness only data set (*WSR*) compared to the similarity only data sets (*MC*, *RG*, *WSS*). This pattern is common among *MSA* and all prior techniques which report on these data sets.

5. A STUDY ON STATISTICAL SIGNIFICANCE

Through the results section, we kept away from declaring state-of-the-art method. That was due two facts. First, the differences between reported correlation scores were very small. Second, the size of the data sets was not that large to accommodate for such small differences. These two facts raise a question about the statistical significance of improvement reported by some method A compared to another well performing method B.

We hypothesize that the best method is not necessarily the one that gives the highest correlation score. In other words, being state-of-the-art does not require giving the highest correlation, rather giving a relatively high score that makes any other higher score statistically insignificant.

To test our hypothesis, we decided to perform statistical significance tests on the top reported correlations. Initially we targeted *Word2Vec*, *GloVe*, *ADW*, and *NASARI* besides *MSA*. We contacted several authors and some of them thankfully provided us with pairwise relatedness scores on corresponding benchmark data sets. We also utilized the publicly available semantic vectors of some models like [Baroni et al., 2014] predict vectors.

To measure statistical significance, we performed Steiger’s Z significance test [Steiger, 1980]. The purpose of this test is to evaluate whether the difference between two dependent correlations obtained from the same sample is statistically significant or not, i.e., whether the two correlations are statistically equivalent.

Steiger’s Z test requires to calculate the correlation between the two

correlations. We applied the tests using reported Spearman correlations (ρ) as it is more commonly used than Pearson (r) correlation. We conducted the tests using correlation scores of *MSA*’s tuned model on *WS* data set, *Word2Vec*, *ADW*, and *NASARI*.

Table 5, shows the results using 1-tailed test with significance level 0.05. For each data set, we report method-method Spearman correlation (ρ) calculated using reported scores in Table 3 and Table 4. We report *p*-value of the test as well.

On *MC* data set, the difference between *MSA* score and all other methods was statistically insignificant. Only *ADW* score was statistically significant compared to *NASARI*. This implies that *MSA* can be considered statistically a top performer on *MC* data set.

On *RG* data set, *MSA* gave significant improvement over *NASARI*. *ADW* score was significantly better than *Word2Vec*, *NASARI*, and *MSA*. Overall, *ADW* can be considered the best on *RG* data set followed by *MSA* and *Word2Vec* (their ρ scores are 0.92, 0.86, and 0.84 respectively).

On *WSS*, though *MSA* achieved the highest score ($\rho=0.77$), no significant improvement was proved. Therefore, the differences between the four methods can be considered statistically insignificant.

On *WSR*, *WS*, and *MEN* data sets, we could obtain pairwise relatedness scores of *Word2Vec* only. The significance test results indicated that, the improvement of *MSA* over *Word2Vec* on *WS* was statistically insignificant (their ρ scores are 0.77 and 0.76 respectively). On the other hand, *MSA* was statistically better than *Word2Vec* on *WSR* data set (their ρ scores are 0.71 and 0.64 respectively), while *Word2Vec* was statistically better than *MSA* on *MEN* data set (their ρ scores are 0.79 and 0.75 respectively).

This comparative study is one of the main contributions of this paper. To our knowledge, this is the first study that addresses evaluating the statistical significance of results across various semantic relatedness methods. Additionally, this study positioned *MSA* as one of state-of-the-art methods for measuring semantic relatedness.

6. CONCLUSION

In this paper, we presented *MSA*, a novel approach for semantic analysis which employs data mining techniques to create conceptual vector representations of text. *MSA* is motivated by inability of prior concept space models to capture implicit relations between

concepts. To this end, *MSA* mines for implicit concept-concept associations through *Wikipedia*'s "See also" link graph.

Intuitively, "See also" links represent related concepts that might complement the conceptual knowledge about a given concept. Furthermore, it is common in most online encyclopedic portals to have a "See also" or "Related Entries" sections opening the door for more conceptual knowledge augmentation using these resources in the future.

Through empirical results, we demonstrated *MSA*'s effectiveness to measure lexical semantic relatedness on benchmark data sets. In absolute figures, *MSA* could consistently produce higher Pearson correlation scores than other explicit concept space models such as *ESA*, *SSA* on all data sets. Additionally, *MSA* could produce higher scores than *ADW* and *NASARI* on four out of five data sets. On another hand, *MSA* scores were higher than predictive models built using neural networks, e.g., *Word2Vec*.

Regarding Spearman correlation, *MSA* produced the highest correlations on two data sets (*WSS* and *WSR*). Results on other data sets were very competitive in absolute figures. Specifically, *MSA* gave higher Spearman correlations than *GloVe* and *Word2Vec* on both *MC* and *RG* data sets. Additionally, *MSA* gave higher correlation score on *MEN* data set than *Skipgram*, *CW*, and *GloVe* neural network based representations.

The results show competitive performance of *MSA* compared to state-of-the-art methods. More importantly, our method produced higher correlation scores than prior explicit semantics methods such as *ESA* and *SSA*. The good performance demonstrates the potential of *MSA* for augmenting the explicit concept space by other semantically related concepts which contribute to understanding the semantics of the given text.

In this paper, we introduced the first comparative study which evaluates the statistical significance of results from across top performing semantic relatedness methods. We used Steiger's Z significance test to evaluate whether reported correlations from two different methods are statistically equivalent even if they are numerically different. We believe this study will help the research community to better evaluate and position state-of-the-art techniques at different application areas. The study proved that, statistically, *MSA* results are either better than or equivalent to state-of-the-art methods on all data sets except *RG* where *ADW* was better, and *MEN* where *Word2Vec* was better.

MSA is a general purpose semantic analysis approach which builds explicit conceptual representations of textual structures. *MSA*'s concept vectors can be easily interpreted by humans. Therefore, *MSA* could be leveraged in many text understanding applications such as semantic search, textual entailment, word sense disambiguation, resolving vocabulary mismatch, concept tracking, technology mappings, and others.

MSA is an efficient technique because it employs an inverted search index to retrieve semantically related concepts to a given text. Additionally, mining for concept(s)-concept(s) association rules is done offline making it scalable to huge amounts of data.

7. REFERENCES

- [Agirre et al., 2009] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.
- [Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM.
- [Banjade et al., 2015] Banjade, R., Maharjan, N., Niraula, N. B., Rus, V., and Gautam, D. (2015). Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In *Computational Linguistics and Intelligent Text Processing*, pages 335–346. Springer.
- [Baroni et al., 2014] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.
- [Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.
- [Bruni et al., 2014] Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res. (JAIR)*, 49:1–47.
- [Budanitsky and Hirst, 2006] Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.
- [Camacho-Collados et al., 2015] Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015). Nasari: a novel approach to a semantically-aware representation of items. In *Proceedings of NAACL*, pages 567–577.
- [Collobert and Weston, 2008] Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.
- [Deerwester et al., 1990] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.
- [Egozi et al., 2011] Egozi, O., Markovitch, S., and Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29(2):8.
- [Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.
- [Finkelstein et al., 2001] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppín, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.
- [Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.
- [Halawi et al., 2012] Halawi, G., Dror, G., Gabrilovich, E., and Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.
- [Harris, 1954] Harris, Z. S. (1954). Distributional structure. *Word*.

- [Hassan and Mihalcea, 2011] Hassan, S. and Mihalcea, R. (2011). Semantic relatedness using salient semantic analysis. In *AAAI*.
- [Hill et al., 2014] Hill, F., Cho, K., Jean, S., Devin, C., and Bengio, Y. (2014). Not all neural embeddings are born equal. *arXiv preprint arXiv:1410.0718*.
- [Hill and Lewicki, 2007] Hill, T. and Lewicki, P. (2007). *Statistics: Methods and Applications*. StatSoft, Inc.
- [Jarmasz and Szpakowicz, 2004] Jarmasz, M. and Szpakowicz, S. (2004). Roget's thesaurus and semantic similarity. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 2003:111.
- [Landauer et al., 1997] Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417. Citeseer.
- [Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- [Miller and Charles, 1991] Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.
- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.
- [Pilehvar and Navigli, 2015] Pilehvar, M. T. and Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.
- [Radinsky et al., 2011] Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.
- [Rubenstein and Goodenough, 1965] Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.
- [Song and Roth, 2014] Song, Y. and Roth, D. (2014). On dataless hierarchical text classification. In *AAAI*, pages 1579–1585.
- [Song and Roth, 2015] Song, Y. and Roth, D. (2015). Unsupervised sparse vector densification for short text similarity. In *Proceedings of NAACL*.
- [Steiger, 1980] Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.
- [Turney et al., 2010] Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.
- [Zesch et al., 2008] Zesch, T., Müller, C., and Gurevych, I. (2008). Using wiktionary for computing semantic relatedness. In *AAAI*, volume 8, pages 861–866.
- [Zwillinger and Kokoska, 1999] Zwillinger, D. and Kokoska, S. (1999). *CRC standard probability and statistics tables and formulae*. CRC.